Real Choices and Hypothetical Choices

by

Glenn W. Harrison†

April 2022

Forthcoming, Stephane Hess and Andrew Daly (eds.)
Handbook of Choice Modeling
(Northampton, MA: Edward Elgar, Second Edition, 2022).

ABSTRACT

There is evidence that behavior changes when individuals make choices over hypothetical scenarios and stakes rather than real scenarios and stakes. What is the nature of this evidence, and how significant is it for different types of inferences? In particular, does it follow that "stated preferences" inferred from hypothetical choices differ all that much from the "preferences revealed" by real choices? What survey methods can better approximate real choices? What econometric methods allow one to pool hypothetical and real choices, when appropriate? Is the bias between real and hypothetical choices transferable from setting to setting?

† Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA. E-mail contact: gharrison@gsu.edu.

The distinction between real choices and hypothetical choices had traditionally been completely ignored or the focal point of intense inter-disciplinary controversy. In some quarters the terminology distinguishes "stated preferences" and "revealed preferences," where the former means preferences revealed by choices when there are no consequences for the decision maker and the latter means preferences revealed when there are consequences for the decision maker. The issues are the same. Does it matter if choices are hypothetical or real? If so, what can be done about it? Have recent efforts to address the issue of hypothetical bias informed the answer these questions?

There are many variants of "choice experiments" in use and the distinction between real and hypothetical choices affects them all. In the context in which the expression is used in this volume, it refers to any situation in which a decision-maker is asked to rank or choose from two or more alternatives *and* where there are several choices to be made in which one or more attributes of the alternatives are varied. In general there are many more attributes than prices that are varied.

There appears to be no *logical* reason to restrict the term "choice experiments" to hypothetical tasks, although that is common in the area of environmental valuation and marketing. The comparison of hypothetical responses and real responses lies at the heart of tests for incentive compatibility, where the expression "real responses" is then a short hand for any task for which the choices of the decision-maker are related in a *salient* manner to real outcomes. Choices may also be rewarded in a *non-salient* manner, such as if someone was paid $10 to complete a survey irrespective of the responses to the survey. Some draw an artificial line between choice tasks in the context of "contingent valuation" and choice tasks in the context of "stated preference." Both types of tasks are relevant, and suffer from hypothetical bias.

In many social policy settings, the connection between hypothetical and real choices may be more probabilistic and tenuous than the crisp experiments that have been the focus of the academic literature. A survey may have some ill-defined "advisory" role in terms of influencing policy, in some

manner that is often maddeningly vague to experimental economists. But there are sometimes good reasons for such ambiguity, such as when it honestly reflects the true state of scientific knowledge or the political and legal process. We know very little about the effects of these types of ill-defined social consequences for incentive compatibility. We therefore focus here on the crisp light of controlled experiments that involve real and transparent consequences, but we also consider how lessons about incentive compatibility drawn from the sharp contrasts of the laboratory can be transferred to more practical settings in which choice studies are applied.

In section 1 the concept of incentive compatibility is reviewed, since it is at the heart of the passion that some have for considering real choices and dismissing hypothetical choice. The practical lesson, however, is that incentive compatibility means more than providing real consequences of the choices that respondents make. The connection between different choices and different consequences has to make it in the best interests of the respondent to respond truthfully.[1] Further, this connection has to be behaviorally transparent and credible, so that the respondent does not start to second-guess the incentive to respond truthfully.

In sections 2 and 3 the importance of making responses incentive compatible is evaluated. The most directly relevant evidence comes from laboratory experiments, where one can crisply compare environments in which the responses are incentive compatible and those where they are not. This distinction has typically been examined by just looking at choices made when the consequences are hypothetical or imagined, and comparing them to choices made when the consequences are real. There is systematic evidence of differences in responses across a wide range of elicitation procedures. The evidence is not universal, and there are some elicitation procedures and contexts in which the problem of incentive compatibility does not appear to be so serious. But there is no "magic bullet" procedure or

[1] Eliciting a truthful response does not mean that the researcher can always directly infer a latent preference or belief from the response, as discussed in section 2.

question-format that reliably produces the same results in hypothetical and real settings.

Section 4 changes gears. The evidence from sections 2 and 3 establishes that there is a problem to be solved: one cannot just assume the problem of incentive compatibility away, at least if one wants to cite the literature in a systematic way. But there are several constructive ways to mitigate hypothetical bias, or correct for it. One is by *ex ante* "instrument calibration," which is the use of controlled experiments with a particular survey population, scenario, and valuation task to identify the best way to ask the question. In effect, this represents the use of experiments to put operationally meaningful teeth in the "focus group" activities that many choice researchers undertake already, at least for large-scale choice studies used for policy or litigation. The other calibration approach is *ex post* the survey, and uses "statistical calibration" procedures to try to correct for any biases in responses. Again, experiments are used to complement the survey, in this case to identify possible differences in hypothetical and real choices that might be systematically correlated with observable characteristics. These statistical methods can then be used to correct for biases, and also to better identify the appropriate standard errors to attach to estimates derived from choice studies.

Section 5 discusses a number of open issues that have been ignored in previous work, and some possible extensions. Section 6 draws conclusions for practical application of a recognition of the difference between hypothetical and real choices. These conclusions might seem harsh, but the objective is to force hypothetical choice researchers to confess to the potential problem they face, and do *something* constructive about it. But arguing for *something* constructive to be done to mitgate hypothetical bias must not be taken as license to do the first thing that pops into one's head. The current practice is simply to quote the literature selectively, which allows the low-level policy applications of the hypothetical choice method to survive casual scrutiny. Higher-level applications are another matter, where the academic, adversarial and policy stakes are substantial enough to force more scrutiny. In those settings the reputation of the hypothetical choice approach, as currently practiced, is frankly appalling. In large part

this might be due to a now-familiar and justifiable source of lack of confidence in (bad) science, the inability to weed out false positives.[2] But that could change quickly if the problem of incentive compatibility is addressed.

1. What Is Incentive Compatibility?

To illustrate the concept of incentive compatibility in relation to choice behavior, we focus initially on voting behavior in referenda, and then turn quickly to more traditional settings for choice experiments. Apart from the popularity of advisory referenda in non-market valuation settings, the context of voting matches the history of thought on these matters. It is then easy to see the implications for choice experiments defined in a non-voting context.

A. Voting

Consider the design of voting mechanisms for referenda that are incentive compatible and non-dictatorial.[3] In the case of voting mechanisms involving the selection of an alternative among k-alternatives, $k \geq 3$, it is well known that no such voting procedure exists.[4] It is, however, easier to devise a voting mechanism involving choice among only two alternatives ($k = 2$) that is incentive compatible. One such voting mechanism is simple majority rule. Typically, incentive compatibility for this mechanism requires, in addition to the restriction to two alternatives, the assumption that individuals perceive that their utilities are affected by the outcome of the vote. Thus, if the voter thinks that his behavior will have some impact on the chance that one or the other alternative will be implemented, and

[2] McElreath and Smaldino [2015] and Smaldino and McElreath [2016].

[3] A dictatorial mechanism is one in which the outcome always reflects the preferences of one specific agent, independent of the preferences of others.

[4] See Gibbard [1973] and Satterthwaite [1975] for the original statements of this theorem, and Moulin [1988] for an exposition.

that his utility will be affected by the outcome, the voter has a *positive* incentive to behave truthfully and vote honestly.

Recent work on institution design using the Revelation Principle employs incentive compatibility as a formal constraint. This formulation uses a much stronger assumption, called Epsilon Truthfulness: *If the agent is indifferent between lying and telling the truth, assume he tells the truth*.[5] It is important that one recognize Epsilon Truthfulness for what it is: an *assertion* or assumption that is regarded by many as excessively strong and that does not enjoy an empirical foundation. It facilitates the proving of theorems, and that is about it. The validity of Epsilon Truthfulness remains an open empirical question.

In the literature concerned with the use of hypothetical choices for valuing environmental goods, the Epsilon Truthfulness assumption is often applied to *hypothetical* referenda. For example, Mitchell and Carson [1989; p.151] state that:

> We also showed that the discrete-choice referendum model was incentive-compatible in the sense that a person could do no better than vote yes if her WTP for a good being valued by this approach was at least as large as the tax price, and to vote no if this was not the case. This finding offers the possibility of framing contingent valuation questions so that they possess theoretically ideal and truthful demand-revelation properties.

Since one cannot know *a priori* whether or not subjects in a choice study will feel that their utilities will be affected by the outcome of a hypothetical vote, such assertions of incentive compatibility require that one *assume* that subjects will behave as they do in real referenda. That is, one invokes a form of the Epsilon Truthfulness assumption.

The question as to whether or not a hypothetical referendum using majority rule is incentive compatible has become an important policy issue given its prominence in proposed guidelines for applications of Contingent Valuation (CV) for estimating environmental damages using stated choice

---

[5] See Rasmussen [1989; p.161]. The Epsilon Truthfulness assumption is used in formal mechanism design problems when the incentive constraints are defined so as to ensure that the expected utility to each agent from a truthful report is greater than *or equal to* the expected utility from any other feasible report.

methods. In proposed rules for using the CV method, both the Department of the Interior (DOI) [1994; p.23102] and the National Oceanographic and Atmospheric Administration (NOAA) [1994; p.1144] assert that, in applications of CV

> ... the voting format is incentive compatible. If respondents desire the program at the stated price, they must reveal their preferences and vote for the program.[6]

This proposed prescription for public policy is based on an assumption that presupposes acceptance of the hypothesis: a voter's behavior is independent of the use of a real or hypothetical referendum mechanism. This hypothesis, and therefore the credibility of the incentive compatibility assumption for hypothetical referenda, has been empirically tested by Cummings, Elliott, Harrison and Murphy [1997].

Our focus here will be on one of the possible reasons for the lack of incentive compatibility of stated choice experiments: hypothetical bias. This bias is said to occur whenever there is a difference between the choices made when the subjects face real consequences from their actions compared to the choices made where they face no real consequences from their actions. However, in many settings of interest to stated choice researchers in environmental economics who deal with public goods, there may be another source deriving from the propensity to free ride on the provision of others. The propensity to free ride[7] has been shown to be alive and well in the laboratory, as the early survey by Ledyard [1995]

---

[6] The adoption of this assertion by the DOI and NOAA is apparently based on a reference to the following statement that appears in an appendix to the NOAA Panel report of Arrow, Solow, Portney, Leamer, Radner and Schuman [1993]: "As already noted, such a question form (a dichotomous choice question posed as a vote for or against a level of taxation) also has advantage in terms of incentive compatibility" (p. 4612). This reference ignores, however, the text of the NOAA Panel's report which includes a lengthy discussion of the advantages and disadvantages of the referendum format used in the *hypothetical* setting of an application of the CV method (pp. 4606-4607), discussions which belie the later assertion of incentive compatibility. Among the disadvantages discussed by them are respondent's reactions to a hypothetical survey, the fact that there can be no real implication that a tax will actually be levied and the damage actually repaired or avoided. Thus, the NOAA Panel suggests that "...considerable efforts should be made to induce respondents to take the question seriously, and that the CV instrument should contain other questions designed to detect whether the respondent has done so" [1993; p.4606]. Further, the NOAA Panel notes a further problem that could detract from the reliability of CV responses: "A feeling that one's vote will have no significant effect on the outcome of the hypothetical referendum, leading to no reply or an unconsidered one...." [1993; p.4607].

[7] Free riding is said to occur when a subject does not make any contribution to the provision of a public good that is valued by the subject.

documented. Harrison and Hirshleifer [1995] also show that it varies theoretically and behaviorally with the nature of the production process used to aggregate private contributions into a public good, such as one finds with threshold effects in many public goods (e.g., health effects of pollutants, species extinction). It is difficult to say *a priori* if free riding bias is greater than the hypothetical bias problem. There is a dearth of studies of the interaction of the two biases.

To answer the question posed at the outset, incentive compatibility will be measured in terms of differences in responses between hypothetical and real environments, *and* where the real environment has been designed to encourage truthful responses. This will normally mean that the scenario is not imaginary, but it is the actual, non-hypothetical consequence that is the behavioral trace that we use to identify deviations from incentive compatibility.

Knowledge that the respondent will answer truthfully normally comes from *a priori* reasoning about rational responses to known incentives. So this is the methodological domain of causal modeling, not mere correlation (McElreath [2020; ch.1]). But we will also want to be cognizant of the need to ensure that the respondent sees what is *a priori* obvious to the (academic) analyst.[8] For example, we prefer mechanisms for which it is a dominant strategy to tell the truth, where this can be explained to the respondent in a non-technical manner, and where the verification of this fact is a simple matter for the subject. Sometimes we cannot have this ideal behavioral environment. Rational responses may be truthful only in some strategic Nash Equilibrium, so the respondent has to make some guess as to the rationality of other players. Or the respondent might not understand the simple explanation given, or suspect the surveyor of deception, in which case "all bets are off" when it comes to claims of incentive compatibility. All of this calls for some theory, or theories, about the processes generating the observed

[8] This point can be stated more formally by thinking of the choice study as a game between the surveyor and the respondent. There is a difference between complete information and common knowledge in strategic games that captures this distinction. Surveyors can tell subjects something that is true, but that is a not the same thing as knowing that subjects believe those things to be true. Linguistics has rich traditions that help us think about the everyday transition to common knowledge in these settings.

data. This is not easy, or attractive in an era of "point and click" statistical computing.

B. Willingness to Pay

In the setting of eliciting willingness to pay (WTP) for some private good or service, there are many mechanisms that are incentive compatible. The simplest is to just ask someone if they are willing to give you $5 for some object, and give it to them if they say yes *and* give you the $5. This is the basis of the Dichotomous Choice (DC) task considered by Cummings, Harrison and Rutström [1995] in simple experiments with a juicer. In the context of auctions, the Vickrey sealed-bid auction is another example: N>1 people bid for the object, the highest bidder receives the object, and she pays the second highest price. The English real-time auction is theoretically isomorphic to the Vickrey auction: the price is called out at $0 and steadily increments in real time, N>1 people sit down when they do not want to pay that price for the object, and literally "the last one standing" gets the object at the price when the second-last person sits down.[9] The Becker, DeGroot and Marschak mechanism is a simulated version of the Vickrey auction: a subject is given the object, states a price they are willing to sell it at, a simulated buying price is generated, and the subject parts with the object if the stated selling price is below the buying price. These alternatives are evaluated by Rutström [1998] in simple experiments with chocolate truffles.

There is a distinction between something being incentive compatible in theory and incentive compatible in terms of behavioral responses. Many subjects just do not understand that it is in their best interests to report their true valuation in response to Vickrey auctions or Becker, DeGroot and Marcshak simulated auctions. When one is not testing if the subject understands that property, it is common to have experimental instructions explain it to the subject. Many studies simply assert that subjects understood this property, and move on. These issues do not arise with binary choice tasks,

[9] A multiple-unit analogue of the Vickrey auction, the Uniform Price auction, is evaluated in experiments by Cox, Smith and Walker [1985].

which have become the staple in many settings, even though one is eliciting minimal information from each choice observation.

C. Telling the Truth and Inferring Latent Constructs

Having a mechanism that gets someone to respond truthfully is one thing, and often enough for inferences about voting preference or WTP to be made. But it is not always, or even normally, enough. Consider, for example, getting someone to report their beliefs about some event. There are well-known scoring rules that provide an incentive for a *risk-neutral* subject to truthfully report her beliefs, whether one is considering binary events or multi-valued events. But what about risk averse subjects? In that case, these (proper) scoring rules still elicit a truthful response, but one has to jointly elicit risk preferences from the subject and undertake some calculations to infer their latent belief (e.g., Andersen, Fountain, Harrison and Rutström [2014] and Harrison, Martínez-Correa, Swarthout and Ulm [2017]). Often one hears researchers say that one must "correct" reported beliefs for risk aversion, but that is conceptually incorrect. The reports are truthful, but what one infers from them depends on theory and appropriate designs.[10]

Another example arises from binary choices over risky lotteries, which are an incentive compatible manner to find out which lottery an individual prefers. But inferring risk preferences from that choice depends on theories of risk preferences and appropriate econometric methods, reviewed by Harrison and Rutström [2008]. In this setting it is tempting for researchers to try to elicit more information than a binary choice, such as the Certainty Equivalent (CE) of a lottery. Armed with the CE, one can then directly infer the Risk Premium (RP) as the Expected Value less the CE. But one must

[10] There exist more complicated elicitation methods that, in theory, allow one to directly infer latent beliefs from reports by "risk neutralizing" the subject's responses. Harrison, Martínez-Correa and Swarthout [2014] and Harrison, Martínez-Correa, Swarthout and Ulm [2015] evaluate these methods for eliciting beliefs over binary and non-binary events, respectively. The challenge, of course, is to have confidence that the subject understands the more complicated task.

still infer risk preferences from the RP, and it does not identify the utility function and/or probability weighting functions the individual might be using.

A final example. Choices over a certain amount of money to be provided at time t or a larger amount of money to be provided at time T, for T >t and t greater than or equal to today, can be truthfully elicited using DC choices (e.g., Coller and Williams [1999] and Harrison, Lau and Williams [2002]). But inferences about latent time preferences do not follow directly from those choice data unless one corrects for non-linearities in utility functions defined over these amounts of money. Joint elicitation of risk and time preferences is one way to infer the true latent time preferences from the true choice data over money: Andersen, Harrison, Lau and Rutström [2008][2014]. And again, it is not that one "corrects" the DC choice data for diminishing marginal utility: one only draws inferences from those correct, true data about the latent time preferences when combining the data with theory and appropriate econometrics.[11]

2. Evidence of Hypothetical Bias from Stylized Choice Tasks

We begin the review of previous evidence by considering the simple cases in which one elicits choices over two alternatives, or where the only attribute that is varied is the cost of the alternative. If we cannot say whether choices are incentive compatible in these settings, we had better give up trying to do so in the more complex settings in which there are more than two alternatives varying in terms of some non-monetary dimension.[12] We simplify things even further by considering elicitation over a private good, for which it is easy to exclude non-purchasers.

[11] And, yet again, there exist more complicated elicitation mechanisms that and have been proposed, that seek to avoid these extra steps involving theory and econometrics: see Andreoni and Sprenger [2012] and Laury, McInnes and Swarthout [2012]. These mechanisms are deeply problematic, for many reasons: *caveat emptor* !

[12] Svenningsen and Jacobsen [2018] is a useful reminder that many of the goods or services we are interested in can have moral attributes for individuals, and that this matters, as it should, for inferences about hypothetical bias.

A DC elicitation in this setting is just a "take it or leave it" offer, much like the posted-offer trading institution studied by experimental economists for many years. As noted earlier, the difference is that the experimenter presents the subjects with a price, and the subject responds "yes" or "no" if she is willing to pay that amount. The subject gets the commodity if and only if they say "yes," and then part with their money. The consequences of a "yes" response are real, and not imagined. Incentive compatibility is apparent, at least in the usual partial-equilibrium settings in which such things are discussed.[13]

Cummings, Harrison and Rutström [1995] (CHR) designed some of the simplest experiments that have probably ever been run, just to expose the emptiness of the claims of those that would simply *assert* that hypothetical responses are the same as real responses in a DC setting. Subjects were randomly assigned to one of two rooms, the only difference being the use of hypothetical or real language in the instructions. An electric juicer was displayed, and passed around the room with the price tag removed

[13] Carson, Flores and Meade [2001; p.191] appear to take issue with this claim, but one simply has to parse what they say carefully to understand it as actually in agreement: "For provision of private or quasi-public goods, a yes response increases the likelihood that the good will be provided, however, the actual decision to purchase the good need not be made until later. Thus, a yes response increases the choice set at no expense." They are not clear on the matter, so one has to fill in the blanks to make sense of this. If the DC involves a real commitment, such that the subject gets the private good if private money is given up, then the yes response does not increase the choice set for free. So they cannot be referring to a real DC response. In the case of a hypothetical DC for private goods, it does not follow that the yes response increases the likelihood of the good being provided. Of course, subjects are entitled to hold whatever false expectations they want, but the explicit script in incentivized choice experiments typically contains nothing intended to lead them to that belief. Carson, Flores and Meade [2001] then suggest how one can make this setting, which can only be interpreted as referring to a hypothetical DC, incentive compatible: "The desirable incentive properties of a binary discrete choice question can be restored in instances where the agent is asked to choose between two alternatives, neither of which represents a strict addition to the choice set." Their footnote 44 then explains what they mean: "It can be shown that what a coercive payment vehicle does is to effectively convert a situation whereby an addition to the choice set (e.g., a new public good) *looks like* a choice between two alternatives, neither of which is a subset of the other, by ensuring the extraction of payment for the good" (emphasis added). So this is just saying that one can make a hypothetical DC incentive compatible by requiring real payment, which is the point that Cummings, Harrison and Rutström [1995] viewed as apparent and hardly in need of notation and proof. The words "look like" are problematic to an experimental economist. They suggest that one must rely on subjects misunderstanding the hypothetical nature of the task in order for it to be incentive compatible. But if subjects misunderstand part of the instructions, how does one know that they have understood all of the rest? Circular "logic" of this kind is precisely why one needs crisp, incentivized experiments.

or blacked-out. The display box for the juicer had some informative blurb about the product, as well as pictures of it "in action." Subjects were asked to say whether or not they would be willing to pay some stated amount for the good.

The hypothetical subjects responded much more positively than the real subjects. Since the private sources funding these experiments did not believe that "students were real people," the subjects were non-student adults drawn from church groups. The same qualitative results were obtained with students, with the same commodity and with different commodities. Comparable results have been obtained in a willingness to accept setting by Nape et al. [2003].

In response to the experimental results of CHR, some proponents of hypothetical surveys argued that their claims for the incentive-compatibility of the DC approach actually pertained to simple majority rule settings in which there was some referendum over just two social choices. Somehow that setting provides the context that subjects need to spot the incentive compatibility, or so it was argued. Again, it is apparent that this context is incentive-compatible if subjects face real consequences.

Cummings, Elliott, Harrison and Murphy (CEHM) [1997] therefore undertook simple majority rule experiments for an actual public good. After earning some income, in addition to their show-up fee, subjects were asked to vote on a proposition that would have each of them contribute a specified amount towards this public good. If the majority said "yes," all had to pay. The key treatments were again the use of hypothetical or real payments, and again there was significant evidence of hypothetical bias.

3. Evidence of Hypothetical Bias from Choice Experiments

We now reconsider more closely the evidence for hypothetical bias from several studies that are closer to the choice modeling environment considered in this volume. Overall, the evidence is that hypothetical bias exists and needs to be worried about: hypothetical choices are not reliably incentive

compatible, even if we live in a world of occasional false positives and false negatives. But there is a glimmer or two of good news, and certain settings in which the extent of hypothetical bias might be minimal. The task is to try to understand this variation in the behavioral extent of the bias, not just document it. Only by understanding it can one design stated choice studies that mitigate it reliably.

A. Multiple Price Lists

A direct extension of the DC choice task is to implicitly offer the subject three choices: buy the good at one stated price, buy the good at another stated price, or keep your money. In this case, known in the experimental literature as an Multiple Price List (MPL) auction, the subject is actually asked to make two choices: say "yes" or "no" to whether the good would be purchased at the first price, and make a similar choice at the second price. The subject can effectively make the third choice by saying "no" to both of these two initial choices. The MPL can be made incentive-compatible by telling the subject that one of the choices will be picked at random for implementation.

The MPL design has been demonstrated to exhibit hypothetical bias in the elicitation of risk attitudes by Holt and Laury [2002)[2005] and Harrison [2005], and in the elicitation of individual discount rates by Coller and Williams [1999].

B. Conjoint Choice Experiments

Conjoint choice tasks involve several choices being posed to subjects, in the spirit of the revealed preference logic. Each choice involves the subject reporting a preference over two or more bundles, where a bundle is defined by a set of characteristics of one or more commodities. The simplest example would be where the commodity is the same in all bundles, but price is the only characteristic varied. This special case is just the MPL discussed above, in which the subject may be constrained to just pick one of the prices (if any). The most popular variant is where price and non-price characteristics

are allowed to vary across the choices. For example, one bundle might be a lower quality version of the good at some lower price, one bundle might be a higher quality version at a higher price, and one bundle is the status quo in which nothing is purchased. The subject might be asked to pick one of these three bundles in one choice task (or to provide a ranking).

Typically there are several such choices. To continue the example, the qualities might be varied and/or the prices on offer varied. By asking the subject to make a series of such choices, and picking one at random for playing out[14], the subjects preferences over the characteristics can be "captured" in the familiar revealed preference manner. Since each choice reflects the preferences of the subject, if one is selected for implementation independently [15] of the subject's responses, the method is obviously incentive-compatible.[16] Furthermore, the incentive to reveal true preferences is relatively transparent.

This set of variants goes by far too many names in the literature. The expression "choice

---

[14] That is, one task is selected after all choices have been made, and the subject plays it out and receives the consequences. This avoids the potentially contaminating effects of changes in real income if one plays out all choices sequentially.

[15] As a procedural matter, experimental economists generally rely on physical randomizing devices, such as die and bingo cages, when randomization plays a central role in the mechanism. There is a long tradition in psychology of subjects second-guessing computer-generated random numbers, and the unfortunate use of deception in many fields from which economists recruit subjects makes it impossible to rely on the subject trusting the experimenter in such things.

[16] The manner in which survey proponents quickly shift ground when confronted by uncomfortable evidence of hypothetical bias is well illustrated by Carson [1997; fn.7): "Once the strategic incentives in the single-private-good case are grasped, it should not be surprising that the marketing research literature evolved away from the single-good case to the multiple-good case, where it is possible to restore some of the incentives for truthful preference revelation." This assertion is hard to understand. There are incentives for truthful revelation if the single DC question for private goods involves real consequences; otherwise, there are simply no incentives without untenable assumptions. The same is true if there are multiple DC questions, providing the real consequences only apply to one of them. Of course, one must temper this formal statement by a modicum of common sense when it comes to the strengths of incentives: Buckell, White and Shang [2020] defend an incentive treatment that gave subjects a 1 in 1,154 chance of facing a real consequence. That is a chance of 0.00086, or 0.086 of a percentage point. One just has to smile at attempts (p. 3) to defend this type of design when it comes to the incentive treatment: "We did not give the respondents the probability precisely because we wanted the value of the incentives to be more salient than the probability of payoff, thereby strengthening the respondents' beliefs that it would be better to report accurately and truthfully." Of course the opposite is true in experimental design: failing to control for a potential confound does not mean you can just explicitly assume it has no effect, even if that is often done implicitly.

experiments" is popular, but too generic to be accurate. A reference to "conjoint analysis" helps differentiate the method, but at the cost of semantic opacity. In the end, the expression "revealed preference methods" serves to describe these methods well, and connect them to a long and honorable tradition in economics since Samuelson [1938], Afriat [1967] and Varian [1982][1983].

Several studies examine hypothetical bias in this revealed preference elicitation method, at least as it is applied to valuation and ranking.

Allocating Money to Environmental Projects

Carlsson and Martinsson [2001] allow subjects to allocate real money to 2 environmental projects, varying 3 characteristics: the amount of money the subject personally receives, the amount of money donated to an environmental project by the researchers, and the specific World Wildlife Fund project that the donation should go to. They conclude that the real and hypothetical response are statistically indistinguishable, using statistical models commonly used in this literature.

However, several problems with their experiment make it hard to draw reliable inferences. First, and most seriously, the real treatments were all in-sample: each subject gave a series of hypothetical responses, and then gave real responses. There are obvious ways to test for order effects in such designs, as used by CHR for example, but they are an obvious confound here. Second, the subjects were allocating "house money" with respect to the donation, rather than their own. This made it hard to implement a status quo decision, since it would have been dominated by the donation options if the subject had even the slightest value for the environmental project. On the other hand, there is a concern that these are all artificial, forced decisions that might not reflect how subjects allocate monies according to their true preferences (unless one makes strong separability assumptions). Third, all three environmental projects were administered by the same organization, which leads the subject to view them as perfect substitutes. This perception is enhanced by a (rational) belief that the organization was

free to re-allocate un-tied funds residually, such that there is no net effect on the specific project. Thus the subjects may well have rationally been indifferent over this characteristic.[17]

Valuing Beef

Lusk and Schroeder [2004] conduct a careful test of hypothetical bias for the valuation of beef using revealed preference methods. They consider 5 different types of steak, and vary the relative prices of each steak type over 17 choices. For the subjects facing a real task, one of the 17 choices was to be selected at random for implementation. Subjects also considered a "none of these" option that allowed them not to purchase any steak. Each steak type was a 12oz steak, and subjects were told that the baseline steak, a "generic steak" with no label, had a market price of $6.07 at a local supermarket. Each subject received a $40 endowment at the outset of the experiment, making payment feasible for those in the real treatment. Applying the statistical methods commonly used to analyze these data, they find significant differences between hypothetical and real responses. Specifically, they find that the marginal values of the attributes between hypothetical and real are identical but that the propensity to purchase, attributes held constant, is higher in the hypothetical case.

More experimental tests of the revealed preference approach are likely. I conjecture that the experimental and statistical treatment of the "no buy" option will be critical to the evaluation of this approach. It is plausible that hypothetical bias will manifest itself in the "buy something" versus "buy nothing" stage in decision-making, and not so much in the "buy this" or "buy that" stage that conditionally follows.[18] Indeed, this hypothesis has been one of the implicit attractions of the method. The idea is that one can then focus on the second stage to ascertain the value placed on characteristics.

[17] When subjects are indifferent over options, it does not follow that they will choose at random. They might use other heuristics to pick choices which exhibit systematic biases. For example, concern with a possible left-right bias leads experimental economists looking at lottery choice behavior to randomize the order of presentation.

[18] See List, Sinha and Taylor [2006] for some evidence consistent with this conjecture.

But this promise may be illusory if one of the characteristics varied is price and separability in decisions is not appropriate. In this case the latent utility specification implies that changes in price spill over from the "buy this or buy that" nest of the utility function and influence the "buy or no-buy" decision.

Ranking Mortality Risks

Harrison and Rutström [2006a] report the results of a conjoint choice ranking experiment in which there was a marked lack of hypothetical bias. Their task involved subjects ranking the 12 major causes of death in the United States. The task was broken down for each subject according to broad age groups. Thus a subject aged 25 was asked to state 12 rankings for deaths in the age group 15 to 24, 12 more rankings for deaths in the age group 25 to 44, 12 more rankings for the age group 45 to 64, and finally 12 rankings for those 65 and over. In the real rewards treatment the subject was simply paid $1 for every correct ranking. Thus the subject could earn up to $48 in the session.

The hypothetical versions of the survey instrument replaced the text in the original versions which described the salient reward for accuracy. The replacement text was very simple:

> You will be paid $10 for your time. We would like you to try to rank these as accurately as you can, compared to the official tabulations put out by the U.S. Department of Health. When you have finished please check that all cells in the table below are filled in.

The experiment was otherwise administered identically to the others with salient rewards, using a between-subjects design. There were 95 subjects in the hypothetical rewards experiments[19] and 45 subjects in the salient rewards experiments. The rank errors for the hypothetical (H) sessions are virtually identical to those in the real (R) sessions. The average rank error in the H sessions is 2.15, compared to 2.00 in the R sessions. Moreover, the standard deviation in the H sessions is 1.95, which is also close to the 1.90 for the R sessions. Although there has been some evidence to suggest that average

[19] After removing subjects that failed to complete the survey in some respect, there are 91 remaining subjects.

H responses *might* be the same as R responses in *some* settings, it is common to see a significantly higher variance in H responses as noted earlier. A regression analysis confirms the conclusion from the raw descriptive statistics, but when appropriate controls are added.

This conclusion from the hypothetical survey variant is a surprise, given the extensive literature on the extent of hypothetical bias: the responses obtained in *this hypothetical setting* are statistically identical to those found in a real setting. The hypothetical setting implemented here should perhaps be better referred to as a non-salient experiment. Subjects were rewarded for participating, with a fixed show-up fee of $10. The hypothetical surveys popular in the field rarely reward subjects for participating, although it has occurred in some cases. There could be a difference between a non-salient experiment and "truly hypothetical" experiments.

One feature of the vast literature on hypothetical bias is that it deals almost exclusively with *valuation* tasks and binary *choice* tasks, rather than *ranking* tasks.[20] The experimental task of Harrison and Rutström [2006a] is a ranking task. It is also possible that the evidence on hypothetical bias in valuation settings simply does not apply so readily to ranking tasks.

This conjecture is worth expanding on, since it suggests some important directions for further research. One account of hypothetical bias that is consistent with these data runs as follows. Assume that subjects come into an experiment task and initially form some beliefs as to the "range of feasible responses," and that they then use some heuristic to "narrow down" a more precise response within that range. It is plausible that hypothetical bias could affect the first step, but not be so important for the second step. If that were the case, then a task that constrained the range of feasible responses, such as our ranking task that restricts the subjects to choose ranks between 1 and 12, might not suffer from hypothetical bias. On the other hand, a valuation task might plausibly elicit extreme responses in a

[20] See Harrison and Rutström [2006b] for one review.

hypothetical setting, as subjects note that they could just as easily say that they would pay nothing as say that they would pay a million dollars. In this setting there is no natural constraint, such as comparing to one's budget, to restrict feasible responses. Hence the second stage of the posited decision process would be applied to different feasible ranges, and even if the second stage were roughly the same for hypothetical and real tasks, if the first stage were sufficiently different then the final response could be very different. This is speculation, of course. The experiment considered here does not provide any evidence for this specific thought process, but it does serve to rationalize the results.

4. Mitigating Hypothetical Bias

There are two broad ways in which one can try to mitigate hypothetical bias: by means of instrument calibration before the survey (trying out different "wordings" to generate less biased hypothetical responses), or by means of statistical calibration after the survey (estimating hypothetical bias functions that can be used to then correct for that bias). Harrison [2006b] surveys these two calibration methods in greater detail.

A. Instrument Calibration

The idea of instrument calibration has already generated two important innovations in the way in which hypothetical questions have been posed: recognition of some uncertainty in the subject's understanding of what a "hypothetical yes" means (Blumenschein et al. [1998)[2001]), and the role of "cheap talk" scripts directly encouraging subjects to avoid hypothetical bias (Cummings, Harrison and Osborne [1995], Cummings and Taylor [1998], List [2001], Aadland and Caplan [2003], Brown, Ajzen and Hrubes [2003], Özdermir, Johnson and Hauber [2009], Jacquemet et al. [2013] and de-Magistris et al. [2013]).

The evidence for these procedures is mixed. Allowing for some uncertainty can allow one to

adjust hypothetical responses to better match real responses, but presumes that one knows *ex ante* what threshold of uncertainty is appropriate to apply. Simply showing that there exists a threshold that can make the hypothetical responses match the real responses, once you look at the hypothetical and real responses, is not particularly useful unless that threshold provides some out-of-sample predictive power. Similarly, the effects of "cheap talk" appear to be context-specific, which simply means that one has to test its effect in each context rather than assume it works in all contexts.

B. Statistical Calibration

The essential idea underlying the statistical calibration approach, developed by Blackburn, Harrison and Rutström [1994], is that a hypothetical survey provides an informative, but statistically biased, indicator of the subject's true willingness to pay for a good or service. The trick is how to estimate and apply such bias functions. They propose doing so with the *complementary* use of field elicitation procedures that use hypothetical surveys, laboratory elicitation procedures that use hypothetical and non-hypothetical surveys, and laboratory elicitation procedures that use incentive-compatible institutions.[21]

Consider the analogy of a watch that is always 10 minutes slow to introduce the idea of a *statistical bias function* for hypothetical surveys. The point of the analogy is that hypothetical responses can still be informative about real responses if the bias between the two is systematic and predictable. The watch that is always 10 minutes slow can be informative, but only if the error is *known* to the decision maker and if it is *transferable* to other instances (i.e., the watch does not get further behind the times over time).

Blackburn, Harrison and Rutström [1994] define a "known bias function" as one that is a

[21] Related work on statistical calibration functions includes Fox et al. [1998], Johannesson et al. [1999] and List and Shogren [1998, 2002].

systematic statistical function of the socio-economic characteristics of the sample. If this bias is not mere noise then one can say that it is "knowable" to a decision maker. They then test if the bias function is transferable to a distinct sample valuing a distinct good, and conclude that it is. In other words, they show that one can use the bias function estimated from one instance to calibrate the hypothetical responses in another instance, and that the *calibrated hypothetical* responses statistically match those observed in a paired *real* elicitation procedure. Johannesson et al. [1999] extend this analysis to consider responses in which subjects report the confidence with which they would hypothetically purchase the good at the stated price, and find that information on that confidence is a valuable predictor of hypothetical bias.

The upshot of the statistical calibration approach is a simple comparison of the original responses to the hypothetical survey and a set of calibrated responses that the same subjects *would have made* if asked to make a real economic commitment in the context of an incentive-compatible procedure. This approach does not predetermine the conclusion that the hypothetical survey is "wrong." If the hypothetical survey is actually eliciting what its proponents say that it is, then the calibration procedure should say so. In this sense, calibration can be seen as a way of validating "good hypothetical surveys" and correcting for the biases of "bad hypothetical surveys."[22]

The statistical calibration approach can do more than simply pointing out the possible bias of a hypothetical choice survey. It can also evaluate the confidence with which one can infer statistics such as the population mean from a given survey. In other words, a decision maker is often interested in the bounds for a valuation that fall within prescribed confidence intervals. Existing hypothetical surveys often convey a false sense of accuracy in this respect. A calibration approach might indicate that the

[22] Mitchell and Carson [1989] provide a popular and detailed review of many of the traits of "bad hypothetical surveys." One might question the importance of some of these traits, but that debate is beyond the scope of this review.

population mean inferred from a hypothetical survey is reliable in the sense of being unbiased, but that the standard deviation was much larger than the hypothetical survey would directly suggest. This type of extra information can be valuable to a risk-averse decision maker.

There have been two variants on this idea of statistical calibration: one from the marketing literature dealing with the pooling of responses from hypothetical and real data process, and one from the experimental literature dealing with in-sample calibration.

Pooling Responses From Different Mechanisms

Building on long-standing approaches in marketing, a different statistical calibration tradition seeks to recover similarities and differences in preferences from data drawn from various institutions. The original objective was "data enrichment," which is a useful way to view the goal of complementing data from one source with information from another source. Indeed, the exercise was always preceded by a careful examination of precisely what one could learn from one data source that could not be learned from another, and those insights were often built into the design. For example, attribute effects tend to be positively correlated in real life: the good fishing holes have many of the positive attributes fishermen want. This makes it hard to tease apart the effects of different attributes, which may be important for policy evaluation. Adroit combination of survey methods can mitigate such problems, as illustrated by Adamowicz, Louviere and Williams [1994].

Relatively few applications of this method have employed laboratory data, such that there is at least one data generating mechanism with known incentive compatibility. One exception is Cameron, Poe, Ethier and Schulze [2002]. They implement 6 different hypothetical surveys, and one actual DC survey. All but one of the hypothetical surveys considered the same environmental good as the actual DC survey; the final hypothetical survey used a "conjoint analysis" approach to identify attributes of the good. Their statistical goal was to see if they could recover the same preferences from each data

generation mechanism, with allowances for statistical differences necessitated by the nature of the separate responses (e.g., some were binary, and some were open-ended). They develop a mixture model, in which each data generation mechanism contributes to the overall likelihood function defined over the latent valuation. Although they conclude that they were generally able to recover the same preferences from most of the elicitation methods, their results depend strikingly on the assumed functional forms. Their actual DC response was only at one price, so the corresponding latent WTP function can only be identified if one is prepared to extrapolate from the hypothetical responses. The upshot is a WTP function for the actual response that has a huge standard error, making it hard to reject the null that it is the "same" as the other WTP functions. The problems are clear when one recognizes that the only direct information obtained is that only 27% of the sample would purchase the environmental good at $6 when asked for real, whereas 45% would purchase the good when asked hypothetically.[23] The only information linking the latent WTP functions is the reported income of respondents, along with a raft of assumptions about functional form.

A popular approach to combining data from different sources has been proposed in the stated choice literature: see Hensher, Louviere and Swait [1999], Louviere, Hensher and Swait [2000; ch. 8, 13] and Hensher, Rose and Greene [2015; ch.19] for reviews. One concern with this approach is that it relies on differences in an unidentified "scale parameter" to implement the calibration. Consider the standard probit model of binary choice, to illustrate. One common interpretation of this model is that it reflects a latent and random utility process in which the individual has some cardinal number for each alternative that can be used to rank alternatives. This latent process is assumed to be composed of a deterministic core and an idiosyncratic error. The "error story" varies from literature to literature,[24] but if

[23] This compares the 0-ACT and 1-PDC treatments, which are as close as possible other than the hypothetical nature of the response elicited.

[24] The stated choice literature refers to unobserved individual idiosyncracies of tastes (e.g., Louviere, Hensher and Swait [2000; p.38]), and the stochastic choice literature also refers to trembles or errors by the individual (e.g., Hey [1995)].

one further assumes that it is normally distributed with zero mean *and unit variance* then one obtains the standard probit specification in which the likelihood contribution of each binary choice observation is the cumulative distribution function of a standard normal random variable evaluated at the deterministic component of the latent process. Rescaling the assumed variance only scales up or down the estimated coefficients, since the contribution to the likelihood function depends only on the cumulative distribution below the deterministic component. In the logit specification a comparable normalization is used, in which the variance is set to $\pi^2/3$. Most of the "data enrichment" literature in marketing assumes that the two data sources have the same deterministic component, but allows the scale parameter to vary. This has nothing to say about calibration, as conceived here.

But an extension of this approach does consider the problem of testing if the deterministic components of the two data sources differ, and this nominally has more to do with calibration. The methods employed here were first proposed by Swait and Louviere [1993], and are discussed in Louviere, Hensher and Swait [2000; §8.4]. They entail estimation of a model based solely on hypothetical responses, and then a separate estimation based solely on real responses. In each case the coefficients on the explanatory variables (e.g., sex, age) conditioning the latent process are allowed to differ, including the intercept on the latent process. Then they propose estimation of a "pooled" model in which there is a dummy variable for the data source. Implicitly the pooled model assumes that the coefficients on the explanatory variables *other than the intercept* are the same for the two data sources.[25] The intercepts implicitly differ, if one thinks of there being one latent process for the hypothetical data and one latent process for the real data. Since the data are pooled, the same implicit normalization of variance is applied to the two data sources. Thus one effectively constrains the variance normalizations

[25] This is particularly clear in the exposition of Louviere, Hensher and Swait [2000; p. 237, 244] since they use the notation α^{RP} and α^{SP} for the intercepts from data sources RP and SP, and a common β for the pooled estimates.

to be the same, but allows the intercept to vary according to the data source. The hypothesis of interest is then tested by means of an appropriate comparison of likelihood values.

In effect, this procedure can test if hypothetical and real responses are affected by covariates in the same manner, but not if they differ conditional on the covariates. Thus if respondents have the same propensity to purchase a good at some price, this method can identify that. But if men and women each have the same elevated propensity to "purchase" when the task is hypothetical, this method will not identify that.[26] And the overall likelihood tests will indicate that the data can be pooled, since the method allows the intercepts to differ across the two data sources. Hence claims in Louviere, Hensher and Swait [2000; ch.13] of widespread "preference regularity" across disparate data sources and elicitation methods should not be used as the basis for dismissing the need to calibrate hypothetical and real responses.[27]

On the other hand, the *tests* of preference regularity from the marketing literature are capable of being applied more generally than the methods of *pooling* preferences from different sources. The specifications considered by Louviere, Hensher and Swait [2000; p. 233-236] clearly admit the possibility of marginal valuations differing across hypothetical and real settings.[28] In fact, it is possible to undertake tests that some coefficients are the same while others are different, illustrated by Louviere, Hensher and Swait [2000; §8.4.2]. This is a clear analogue to some parameters in a real/hypothetical experiment being similar (e.g. some marginal effects) but others being quite different (e.g. purchase intention), as illustrated by Lusk and Schroeder [2004]. The appropriate pooling procedures then allow some coefficients to be estimated jointly while others are estimated separately, although there is an obvious

---

[26] Interactions may or may not be identified, but they only complicate the already-complicated picture.

[27] Despite this negative assessment of the potential of this approach for constructive calibration of differences between hypothetical and real responses, the "data enrichment" metaphor that originally motivated this work in marketing is an important and fundamental one for economics.

[28] Louviere, Hensher and Swait [2000; p. 233] use the notation α^{RP} and α^{SP} for the intercepts from data sources RP and SP, and β^{RP} and β^{SP} for the coefficient estimates.

concern with such specification tests leading to reported standard errors that understate the uncertainty over model specification.

Calibrating Responses Within-Sample

Fox et al. [1998] and List and Shogren [1998, 2002] propose a method of calibration which uses hypothetical and real responses from the same subjects for the *same good*.[29] But if one is able to elicit values in a non-hypothetical manner, then why bother in the first place eliciting hypothetical responses that one has to calibrate? The answer is that the relative cost of collecting data may be very different in some settings. It is possible in marketing settings to construct a limited number of "mock ups" of the potential product to be taken to market, but these are often expensive to build due to the lack of scale economies. Similarly, one could imagine in the environmental policy setting that one could actually implement policies on a small scale at some reasonable expense, but that it is prohibitive to do so more widely without some sense of aggregate WTP for the wider project. The local implementation could then be used as the basis for developing (Bayesian) priors as to how one must adjust hypothetical responses for the wider implementation.

These considerations aside, the remaining substantive challenge for calibration is to demonstrate feasibility and utility for the situation of most interest in stated choice valuation, when the underlying target good or project is non-deliverable and one must by definition consider cross-commodity calibration. Again, the work that needs to be done is to better understand when statistical calibration

[29] Fox et al. [1998; p.456] offer two criticisms of the earlier calibration approach of Blackburn et al. [1994]. The first is that it is "inconclusive" since one of the bias functions has relatively large standard errors. But such information on the imprecision of valuations is just as important as information on the point estimates if it correctly conveys the uncertainty of the elicitation process. In other words, it is informative to convey one's imprecision in value estimation if the decision-maker is not neutral to risk. The second criticism is that Blackburn et al. [1994] only elicit a calibration function for one price on a demand schedule in their illustration of their method, and that the calibration function might differ for different prices. This is certainly correct, but hardly a fundamental criticism of the method in general.

works and why, not to just document on occasional "success here" or "failure there." The literature is replete with selective citations to studies that support one position or another; the greater challenge is to explain this disparity in terms of operationally meaningful hypotheses, rather than claim generality for the occasional false positive.[30]

5. Open Issues and Extensions

A. Advisory Referenda and Realism

One feature of hypothetical choice surveys in the field is not well captured by most experiments: the chance that the subject's hypothetical response might influence policy or the level of damages in a lawsuit. To the extent that we are dealing with a subjective belief, such things are intrinsically difficult to control perfectly. In some field surveys, however, there is a deliberate use of explicit language which invites the subject to view their responses as having some chance of affecting real decisions.

If one accepts that field surveys are successful in encouraging *some* subjects to take the survey for real in a subjectively probabilistic sense, then the natural question to ask is: "how realistic does the survey have to be, in the eyes of respondents, before they respond *as if* it were actually real?" In other words, if one can encourage respondents to think that there is some chance that their responses will have an impact, at what point do the subjects behave the way they do in a completely real survey? Obviously this question is well-posed, since we know by construction that they must do so when the chance of the survey being real is 100%. The interesting empirical question is whether any smaller chance of the survey being real will suffice. This question takes on some significance if one can show that the subject will respond realistically even when the chance of the payment and provision being real

[30] There is also a semantic or linguistic confusion between use money as a *numeraire* when eliciting hypothetical choices over non-monetary alternatives *versus* using money as a *payment mode* when eliciting hypothetical choices over non-monetary alternatives (e.g., Vondolia and Navrud [2019]). This has nothing to do with hypothetical bias, since both sets of choices are hypothetical and there is nothing to measure bias against.

is small.

Harrison [2006a] reviews evidence to show that just making surveys "realistic" is not the panacea for hypothetical bias that one might hope.

B. Salient Rewards

Experimental economics differentiates between non-salient rewards and salient rewards. The former refer to rewards that do not vary with performance in the task: for example, an initial endowment of cash, or perhaps the show-up fee.[31] The latter refer to rewards that vary with performance in the task. In parallel to the distinction between fixed and variable costs, these might be called fixed rewards and variable rewards. The hypothetical setting for virtually all of the experiments considered here should be better referred to as an experiment with non-salient rewards, since subjects were typically rewarded for participating. The hypothetical surveys popular in the field rarely reward subjects for participating with a fixed reward, although it has occurred in some cases. There could be a difference between the non-salient experiments which are called "hypothetical" and "truly hypothetical" experiments in which there are no rewards (salient or non-salient). More systematic variation in the non-salient rewards provided in hypothetical choice studies would allow examination of these effects.[32]

[31] The show-up fee is fixed conditional on the subject turning up and participating. It is definitely presumed to be salient with respect to the participation decision.

[32] A conjecture. If subjects are brought in and given a substantial non-salient reward for participating, and given certain "(not so) cheap talk," would they behave as if facing salient rewards? The "(not so) cheap talk" would be something along these lines: "we have given you a large fee for just filling out this hypothetical survey because we value your responses. We are unable to make this a survey with real consequences. But we would like you to consider your responses as if it were real. We are giving you this large fee to encourage you to do that, because we value your careful consideration." The rationale for this treatment is that the payment might set up a "social contract" between the experimenter and subject, leading to a "gift exchange" of cognitive effort in return for the fixed participation fee. The quotation marks flag our fears as to what might happen, but these are easy things to test behaviorally.

C. A Common Defense

One common defense for ignoring hypothetical bias is that an influential survey by Camerer and Hogarth [1999] is casually cited as concluding that there is no evidence of hypothetical bias in simple risky lottery choices. What Camerer and Hogarth [1999] conclude, quite clearly, is that the use of hypothetical rewards makes a difference to the choices observed, but that it does not generally change the inference that they draw about the validity of a particular model oif risk preferences, Expected Utility Theory (EUT). Since tests of EUT typically involve paired comparisons of response rates in two lottery pairs, it is logically possible for there to be (i) differences in choice probabilities in a given lottery depending on whether one use hypothetical or real responses, and (ii) no difference between the effect of the EUT treatment on lottery pair responses rates depending on whether one uses hypothetical or real responses.

Furthermore. Camerer and Hogarth [1999] explicitly exclude from their analysis the mountain of data from experiments on valuation[33] that show hypothetical bias. Their rationale for this exclusion was that economic theory did not provide any guidance as to which set of responses was valid. This is an odd rationale, since there is a well-articulated methodology in experimental economics that is quite precise about the motivational role of salient financial incentives (Smith [1982]). In addition, the experimental literature has generally been careful to consider elicitation mechanisms that provide dominant strategy incentives for honest revelation of valuations, and indeed in most instances explain this to subjects since it is not being tested. Thus economic theory clearly points to the real responses as having a stronger claim to represent true valuations. In any event, the mere fact that hypothetical and real valuations differ so much tells us that at least one of them is wrong! Thus one does not actually need to identify one as reflecting true preferences, even if that is an easy task *a priori*, in order to

[33] The term "valuation" subsumes open-ended elicitation procedures, as well as DC, binary referenda and stated choice tasks.

recognize that there are *differences* in behavior between hypothetical and real choices.

D. Administrative Data

One attractive way to evaluate the possible bias of hypothetical measuring instruments is to compare them to data on real choices that are collected in an administrative capacity. Typically this refers to data collected by government agencies, directly or indirectly. In some countries, such as Denmark, Sweden, Norway and Canada, these data can be accessed by accredited researchers and even linked to auxiliary data sources. And those auxiliary data sources can be hypothetical surveys or or incentivized experiments developed by the researcher.[34]

One limitation of the pairing of these data can matter for inferences about hypothetical bias, but must be taken with a pinch of salt. The data-generating processes behind administratively collected choice data may not match those of the hypothetical choice data. This is more than just the ability to consider combinations of product or service attributed in hypothetical choice settings that have never been observed or considered in actual data. That ability, of course, is one potential strength of hypothetical choice data.[35] Instead, we often do not know the strength of the incentives that individuals faced when making the choices that go into administrative data, since they depend on latent opportunity costs that are hard to measure. One of the points of collecting real choice data in experiments is that one can control the direct monetary (or non-monetary) consequences of one choice over another. To be sure, opportunity costs may still play a role, as they do with surveys. I despise the time needed to take surveys, for example, and react aggressively to them when I perceive attempts to trick me into revealing

[34] For example, one can collect data on individual financial wealth, to evaluate the extent to which the possible earnings from experiments used to elicit risk preferences are integrated with that wealth (e.g., Andersen et al. [2018]). These data can also be used to collect data on individuals that do not participate in surveys or experiments, permitting rich econometric evaluation of the potential effects of sample selection on unobservables (e.g., Harrison, Lau and Yoo [2020]).

[35] Chavez et al. [2020] explore the implications of considering attributes that do not exist for the design, ethics and inferences one draws from (stated and incentivized) choice experiments.

how consistent my choices are (e.g., repeated choices after filler tasks, or repeated choices with reversed response scales). But in an important sense, administratively collected data obviously have great currency. The ideal would be to have data collected in hypothetical surveys *and* incentivized experiments that are as close to each other as possible apart from the obvious difference, and then to link both to comparable administrative data.

In some settings, particularly in transport economics, it has been possible to get the best of both worlds here by collecting data on actual travel choices in an administrative manner, using Global Positioning Systems (GPS) devices. There are valuable comparisons to be made when these are paired, usually for the same subjects, with hypothetical surveys to collect choices over these travel options.

One of the earliest such studies by Nielsen [2004] involved 400 individuals and their cars being fitted with GPS units in Copenhagen.[36] Various pricing schemes were offered during a treatment period, which came before or after a control period that had no such schemes in place. In one strata the two periods were 8 weeks long, and the other strata they were 10 weeks long. Apart from general surveys before and after the GPS field experiment, a stated preference survey was conducted at the outset to infer value of time and response to pricing schemes similar to those actually implemented. There were significant technological issues with the GPS units, but one of the striking results was that the field effects of pricing was much larger the longer the time allowed for the effect. This, of course, is a familiar story about long-run price elasticities being larger than short-run price elasticities, as other inputs to the "family driving production function" became variable rather than fixed. It also appeared that the stated preference survey did not have a time dimension on the responses, which of course mattered for the observed choices. This is not a methodological flaw, so much as an incompletely specified survey.

A comparable design, with much more control, was undertaken in Sydney and described by

[36] An additional 100 subjects were recruited, based on the initial findings, and a very different incentive system used.

Fifer et al. [2010][2014]. The design of the experiment clearly had, as one goal, a controlled comparison of stated preference choice tasks, and observed driving behavior in a GPS-monitored field experiment for 10 weeks. The relevant attributes of the tasks, locations, drives and time frames were comparable. Contemporary modeling procedures for stated preference surveys were used, to allow in for heterogeneity in a flexible manner following the methods reviewed by Hensher, Rose and Greene [2015]. The conclusion [2014; p. 176] was clear: "This research supports the existence of hypothetical bias in [Stated Choice] methods irrespective of the model outcomes used to measure the bias, the rules used to define the bias and the mitigation techniques applied to reduce the bias."

E. Process Data

It has been popular to develop methods to evaluate the decision-making processes that individuals exhibit when making hypothetical and real choices, and to try to detect similarities and differences as a clue as to why they might be different. To take the simplest, and perhaps least interesting, example: what if respondents to hypothetical surveys take less than a second to make complex choice tradeoffs, but respondents to incentivized choices take a minute or two to make otherwise comparable choice tradeoffs? At some *a priori* level one might think that more time must indicate a better quality decision in some sense, but it is the "some sense" that is hard to turn into anything that might be descriptively or normatively rigorous. The fact that time response data is often easy to collect along the way, with computerized response interfaces, does not justify giving it more attention in analyses.[37]

Nonetheless, valuable insights into the decision-making process can be gained by documenting

[37] Bonsall and Lythgoe [2010] illustrate the use of time taken to make hypothetical judgments in stated choice tasks, primarily to infer correlates with self-reported confidence in the response. A remarkable set of disciplines seems to have opinions on hypothetical bias and how it should be conceptualized and measured: Haghani et al. [2021a] is a useful survey of these outer reaches of scholarship.

more about the cognitive steps involved. In economics, eye-trackers have been used to better understand the choice attributes in risky lotteries that are literally looked at more than others (e.g., Harrison and Swarthout [2019]). Data of this kind could be used to evaluate some of the hueristics proposed to evaluate behavior patterns in stated choice settings, and whether they are an artefact of consequences being hypothetical. One excellent example in this respect is the use of a "reference choice" in stated choice tasks, reflecting actual (albeit self-reported) purchasing experiences: see Hess, Rose and Hensher [2008] and below. As another example, consider the heuristic evaluated by Moser and Raffaelli [2014], which is a counterpart to the notion of "similarity relations" from cognitive psychology: the idea that individuals might not differentiate certain attribute levels.

Many of the the mitigation approaches proposed have little or no causal basis in economics, but might provide insights into cognitive processes that could be incorporated into rigorous models. Haghani et al. [2021b; p. 1] offer a dizzying review of speculative mitigation methods that have been floated in recent years:

> Ex-ante bias mitigation methods include cheap talk, real talk, consequentiality scripts, solemn oath scripts, opt-out reminders, budget reminders, honesty priming, induced truth telling, indirect questioning, time to think and pivot designs. Ex-post methods include follow-up certainty calibration scales, respondent perceived consequentiality scales, and revealed-preference assisted estimation.

One can only hope that some of these get evaluated in common settings, with credible metrics for evaluating the extent of any mitigation of hypothetical bias, so that one has can weed out false positives before they take root in policy debates. Haghani et al. [2021b; p. 1] correctly observe that "variation in operational definitions of [hypothetical bias] has prohibited consistent measurement of [hypothetical bias] in [choice experiments]."

F. Bayesian Methods

All of the attempts at *ex post* statistical correction for the possibility of hypothetical bias seem to

have been designed for an era in which one could flexibly and rigorously apply prior beliefs to observed data using Bayesian methods. Rather then search for some scalar, such as the number "3" that pops up in the meta-analyses of WTP by List and Gallet [2001] and Murphy et al. [2005], we should be searching for informed priors about variations in the extent of hypothetical bias from individual to individual. In the spirit of Buckell and Hess [2019] and Coote, Swait and Adamowicz [2021], for example, we should be looking for latent characteistics of preference functions that allow informed statistical calibration of hypothetical and real choices.

In turn, this type of statistical calibration calls for Hierarchical Bayesian Models, where pooled data from a sub-sample of a population can be used to infer predictive posterior beliefs about hypothetical bias on the basis of informed priors.[38] The sub-sample can be given one or other experimental task over private or public goods that can be credibly delivered, and the usual array of observable covariates used to condition pooled estimates of hypothetical bias that can then be combined with the covariates and hypothetical responses of a wider sample from the population to infer calibrated responses if the task had been incentivized. The upshot will be random: some distribution showing the extent of possible biases and the weight we should attach to them. For some individuals the variance of the distribution might be narrow, and for some it might be wide. For some individuals the average of the distribution might be close to zero, for others it could be very different from zero. One can then make informed claims to juries or policy-makers about the credibility that can be attached to different WTP or WTA statements based on hypothetical survey choices.

This approach is "data based" solely in the Bayesian method. Underlying the informed priors are simple experimental tasks that are easy to explain to subjects and also to anyone that has to draw

[38] There are now many introductions to such models in economics, psychology and marketing. See Gao, Harrison and Tchernis [2022] and Rossi, Allenby and McCulloch [2005] for applications in economics and marketing, respectively, each with extensive historical references.

inferences based on them. There is no need for a "general theory of hypothetical bias," such as called for by Loomis [2011; §5].

G. Bias and Confidence

Extensive use of the expression "hypothetical bias" might lead some to focus too much on whether the *average* response from hypothetical choice tasks is the same as the *average* response from incentivized choice tasks. This confuses bias defined in terms of the confidence we might have about difference between two *summary statistics* of two distributions with bias defined in terms of differences in the two distributions as a whole. Even if the averages are the same, it is important to know if the variances and skew of the distributions are the same before one can say that "hypothetical bias" is absent. There is some evidence from controlled experiments that lower incentives lead to greater variability of responses, whether or not there is an effect on the average: see Harrison [1989][1992].

H. Pivot Designs and Hypothetical Scenarios

An important development, latent in many consulting studies using choice experiments and some published studies such as Brownstone and Small [2005], is the use of a "reference point" in the choice set that corresponds to an observed choice by the subject. Set aside for the moment that this "observed choice" is still one that is self-reported by the subject. It could be that the subject just reported the route taken every day for a period when going to and from work, and the researcher then fleshes that out by stating the attributes of that route in terms of typical time, congestion and other characteristics. The idea is then to present this as one of the alternatives to the subject, along with constructed alternatives that are completely hypothetical[39] in the usual sense: see Hess, Rose and

[39] One assumes that logically or physically infeasible combinations are excluded *a priori*, as they usually are.

Hensher [2008] and Hensher, Rose and Greene [2015; §19.6.4]. There is some evidence that hypothetical responses vary when such reference points are included, and that they vary asymmetrically around that reference point.[40]

These designs point to potential issues with *hypothetical scenario construction* as distinct from hypothetical bias in terms of the consequences of the choice being hypothetical or real. Of course, one reason for hypothetical bias could well be rejection of a hypothetical scenario, and this is a serious issue in the contingent valuation context. So it may be useful to consider in more detail "what could possibly go wrong" when stating a hypothetical scenario when it comes to working out what it is that the subject is actually responding to.

One of the first "cultural" differences that strikes an experimental economist dipping his toes into the sea of contingent valuation and stated choice studies is how careful those studies are in their choice of language on some matters and how appallingly vague they are on other matters. The best CV studies spend a lot time, and money, on "focus groups" in which they tinker with minute details of the scenario and the granular resolution of pictures used in displays. But they often leave the most basic of the "rules of the game" for the subject unclear.

For example, consider the words used to describe the scenario in the landmark *Exxon Valdez* oil spill study by Carson, Mitchell, Hanemann, Kopp, Presser and Ruud [1992], undertaken in support of litigation by the Attorney-General of the State of Alaska. Forget the simple majority-rule referendum interpretation used by the researchers, and focus on the words actually presented to the subjects. The relevant passages concerning the provision rule are quite vague.

How might the subjects be interpreting specific passages? Consider one hypothetical subject. He

[40] Loose references to prospect theory to motivate this use of reference points, and the possibility of asymmetric responses, misses their deeper contribution in helping the subject understand the choice context better. The *rigorous* laboratory evidence for loss aversion and prospect theory is just pitiful: see Harrison and Swarthout [2022].

is first told, "In order to prevent damages to the area's natural environment from *another* spill, a special safety program has been proposed. We are conducting this survey to find out whether this special program is worth anything to your household." (p.52). Are the proposers of this program going to provide it no matter what I say, and then come for a contribution afterwards? In this case I should free-ride, even if I value the good. Or are they actually going to use our responses to decide on the program? If so, am I that Mystical Measure-Zero Median voter whose response might "pivot" the whole project into implementation? In this case I should tell the truth.

Actually, the subject just needs to attach some positive subjective probability to the chance of being the decisive voter. As that probability declines, so does the (hypothetical) incentive to tell the truth. So, to paraphrase Dirty Harry the interviewer, "do you feel like a specific order statistic today, punk?" Tough question, and presumably one that the subject has guessed at an answer to. I am just adding additional layers of guesswork to the main story, to make clear the extent of the potential ambiguity involved.

Returning to the script, the subjects are later told, "If the program was approved, here is how it would be paid for." But who will decide if it is to be approved? Me, or is that out of my hands as a respondent? As noted above, the answer matters for my rational response. The subjects *were* asked if they had any questions about how the program would be paid for (p. 55), and had any confusions clarified then. But this is no substitute for the control of being explicit and clear in the prepared part of the survey instrument.

Later in the survey the subjects are told, "Because everyone would bear *part* of the cost, we are using this survey to ask people how they would vote if they had the chance to vote on the program." (p.55). OK, this suggests that the provision rule would be just like those local public school bond issues I always vote on, so the program will (hypothetically) go ahead if more than 50% of those that vote say

"yes" at the price they are asking me to pay.[41] But I am bothered by that phrase "*if* they had the chance to vote": does this mean that they are not actually going to ask me to vote and decide if the program goes ahead, but are just floating the idea to see if I would be willing to pay something for it *after* they go ahead with the program? Again, the basic issue of the provision rule is left unclear. The final statement of relevance does nothing to resolve this possible confusion: "*If* the program cost your household a total of $(amount) would you vote for the program or against it?" (p.56).

Is this just "semantics"? Yes, but it is not "just semantics." Semantics *are* relevant since it is the study of what words mean and how these meanings combine in sentences to form sentence meanings. Semantics, along with syntax and context, are critical determinants of any claim that a sentence in a CV instrument can be unambiguously interpreted. The fact that a unique set of words can have multiple, valid interpretations is well-known in general to CV researchers. Nonetheless, it appears to have also been well-forgotten in this instance, since the subject simply cannot know the rules of the voting game he is being asked to play.

More seriously, *we* cannot claim as outside observers of his survey response that *we know* what the subject is guessing at.[42] We can, of course, guess at what the subject is guessing at. This is what Carson et al. [1992] do when they choose to interpret the responses in one way rather than another, but this is still just a dressed-up guess. Moreover, it is a serious one for the claim that subjects may have an incentive to free ride, quite aside from the hypothetical bias problem.

The general point is that one can avoid *these* problems with more explicit language about the exact conditions under which the program would be implemented and payments elicited. I fear that CV

---

[41] Each household was given a "price" which suggested that others may pay a different "price." This is standard in such referendum formats, and could be due to the vote being on some fixed formula that taxes the household according to assessed wealth. Although the survey does not clarify this for the subjects, it would be an easy matter to do so.

[42] Statistical approaches to the linguistic issue of how people resolve ambiguous sentences in natural languages are becoming quite standard. See, for example, Allen [1995; Ch.7, 10] and the references cited there.

researchers would shy away from such language since it would likely expose to the subject the truth about the hypothetical nature of the survey instrument. The illusory attraction of the frying pan again.

I. Replication

Much of the empirical literature on hypothetical bias comes from an era before it became common to document data and computer code for replication. Without naming names, it is unfortunate that many of the major, recent studies on hypothetical bias, particularly those involving sophisticated econometric methods, do not provide access to data and code. Data privacy is understandable in some cases, but it is common in some fields to see randomized versions of confidential data provided, to allow others to see the details of implementations. One hopes that standards of documenting data and code become more common, now that the logistical costs of doing so have become low.

6. Conclusions

There is no reliable way to trick subjects into thinking that something is in their best interests when it is not. Nonetheless, the literature on hypothetical choice is littered with assertions that one can somehow trick people into believing something that is not true. One probably can, if deception is allowed, but such devices cannot be reliable more than once. The claims tend to take the form, "if we frame the hypothetical task the same way as some real-world task that is incentive compatible, people will view it as incentive compatible." The same view tends to arise in the stated choice literature, but is just a variant on a refrain that has a longer history.

There are some specifications which do appear to mitigate hypothetical bias in some settings, but such instances do not provide a general behavioral proof that can be used as a crutch in other instances. For example, there is *some* evidence that one can isolate hypothetical bias to the "buy or no-buy" stage of a nested purchase decision, and thereby mitigate the effects on demand for a specific

product. Similarly, there is *some* evidence that one can avoid hypothetical bias by using ranking tasks rather than choice or valuation tasks. In each case there are interesting conjectures about the latent decision-making process that provide some basis for believing that the specific results might generalize. But we simply do not know yet, and the danger of generalizing is both obvious and habitually neglected in the stated choice literature. These possibilities should be explored, and evaluated in other settings, before relied on casually to justify avoiding the issue.

The only recommendation that can be made from experiments designed to test for incentive compatibility and hypothetical bias is that one has to address the issue head on. If one can deliver the commodity, which is the case in many stated choice applications in marketing, do so. If it is expensive, such as a beta product, then do so for a sub-sample to check for hypothetical bias and correct it statistically. If it is prohibitive or impossible, which is the case in many stated choice applications in environmental and transportation economics, use controlled experiments for a surrogate good as a complementary tool. That is, find some deliverable private or public good that has some of the attributes of the target good, conduct experiments to measure hypothetical bias using samples drawn from the same population, and use the results to calibrate the instrument and/or the responses using appropriate Bayesian methods. And explore the task specifications that appear to mitigate hypothetical bias. Above all, read with great suspicion any study that casually sweeps the problem under the rug.

References

Aadland, David, and Caplan, Arthur J., "Willingness to Pay for Curbside Recycling with Detection and Mitigation of Hypothetical Bias," *American Journal of Agricultural Economics*, 85, 2003, 492-502.

Adamowicz, Wiktor L., Louviere, Jordan J., and Williams, Michael, "Combining revealed and stated preference methods for valuing environmental amenities," *Journal of Environmental Economics and Management*, 26(3), 1994, 271-292.

Afriat, Sidney, "The Construction of a Utility Function from Expenditure Data," *International Economic Review*, 8, 1967, 67-77.

Andersen, Steffen; Fountain, John; Harrison, Glenn W., and Rutström, E. Elisabet, "Estimating Subjective Probabilities," *Journal of Risk & Uncertainty*, 48, 2014, 207-229.

Andersen, Steffen; Cox, James C.; Harrison, Glenn W.; Lau, Morten I.; Rutström, E. Elisabet, and Sadiraj, Vjollca, "Asset Integration and Attitudes Toward Risk: Theory and Evidence," *Review of Economics and Statistics*, December 2018, 100(5): 816–830.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), 2008, 583-618.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Discounting Behavior: A Reconsideration," *European Economic Review*, 71(1), 2014, 15-33.

Andreoni, James, and Sprenger, Charles, "Estimating Time Preferences from Convex Budgets," *American Economic Review*, 102(7), December 2012, 3333-3356.

Arrow, Kenneth; Solow, Robert; Portney, Paul; Leamer, Edward E.; Radner, Roy; and Schuman, Howard, "Report of the NOAA Panel on Contingent Valuation," *Federal Register*, 58(10), January 15, 1993, 4602-4614.

Blackburn, McKinley; Harrison, Glenn W., and Rutström, E. Elisabet, "Statistical Bias Functions and Informative Hypothetical Surveys," *American Journal of Agricultural Economics*, 76(5), December 1994, 1084-1088.

Blumenschein, Karen; Johannesson, Magnus; Blomquist, Glenn C.; and Liljas, Bengt, and O'Conor, Richard M., "Experimental results on expressed certainty and hypothetical bias in contingent valuation," *Southern Economic Journal*, 65(1), 1998, 169-177.

Blumenschein, Karen; Johannesson, Magnus, and Yokoyama, K., "Hypothetical vs. Real Willingness to Pay in the Health Sector: Results from a Field Experiment," *Journal of Health Economics*, 20(3), May 2001, 441-457.

Bonsall, Peter, and Lythgoe, Bill, "Factors affecting the amount of effort expended in responding to questions in behavioural choice experiments," *Journal of Choice Modelling*, 2[2), 2009, 216-236.

Brown, Thomas C.; Ajzen, Icek, and Hrubes, Daniel, "Further Tests of Entreaties to Avoid Hypothetical Bias in Referendum Contingent Valuation," *Journal of Environmental Economics and Management*, 46[2), September 2003, 353-361.

Brownstone, David, and Small, Kenneth A.,"Valuing Time and Reliability: Assessing the Evidence from Road Pricing Demonstrations," *Transportation Research Part A*, 39(4), 2005, 279-293.

Buckell, John; White, Justin S.; and Shang, Ce, "Can incentive-compatibility reduce hypothetical bias in smokers' experimental choice behavior? A randomized discrete choice experiment," *Journal of Choice Modelling*, 37, December 2020, 100255.

Buckell, John, and Hess, Stephane, "Stubbing out hypothetical bias: improving tobacco market predictions by combining stated and revealed preference data," *Journal of Health Economics*, 65, May 2019, 93-102.

Camerer, Colin F., and Hogarth, Robin M, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 19, December 1999, 7-42.

Cameron, Trudy Ann; Poe, Gregory L.; Ethier, Robert G., and Schulze, William D., "Alternative Non-market Value-Elicitation Methods: Are the Underlying Preferences the Same?" *Journal of Environmental Economics and Management*, 44, 2002, 391-425.

Carlsson, Fredrick, and Martinsson, Peter, "Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments?" *Journal of Environmental Economics and Management*, 41, 2001, 179-192.

Carson, Richard T., "Contingent Valuation: Theoretical Advances and Empirical Tests Since the NOAA Panel," *American Journal of Agricultural Economics*, 79(5), December 1997, 1501-1507.

Carson, Richard T.; Flores, Nicholas E., and Meade, Norman F., "Contingent Valuation: Controversies and Evidence," *Environmental and Resource Economics*, 19, 2001, 173-210.

Carson, Richard T.; Mitchell, Robert C.; Hanemann, W. Michael; Kopp, Raymond J.; Presser, Stanley; and Ruud, Paul A., *A Contingent Valuation Study of Lost Passive Use Values Resulting From the Exxon Valdez Oil Spill* (Anchorage: Attorney General of the State of Alaska, November 1992).

Chavez, Daniel E.; Palma, Marco A.; Nayga Jr., Rodolfo M.; and Mjelde, James W., "Product availability in discrete choice experiments with private goods," *Journal of Choice Modelling*, 36, September 2020, 100225.

Coller, Maribeth, and Williams, Melonie B., "Eliciting Individual Discount Rates," *Experimental Economics*, 2, 1999, 107-127.

Coote, Leonard V.; Swait, Joffre; and Adamowicz, Wiktor L., "Separating Generalizable from Source-specific Preference Heterogeneity in the Fusion of Revealed and Stated Preferences," *Journal of Choice Modelling*, 40, September 2021, 100302.

Cox, James C.; Smith, Vernon L., and Walker, James M., "Expected Revenue in Discriminative and Uniform Price Sealed-Bid Auctions," in V.L. Smith (ed.), *Research in Experimental Economics* (Greenwich, CT: JAI Press, Volume 3, 1985).

Cummings, Ronald G.; Elliott, Steven; Harrison, Glenn W., and Murphy, James, "Are Hypothetical Referenda Incentive Compatible?" *Journal of Political Economy*, 105(3), June 1997, 609-621.

Cummings, Ronald G.; Harrison, Glenn W., and Osborne, Laura L., "Can the Bias of Contingent Valuation Be Reduced? Evidence from the Laboratory," *Economics Working Paper B-95-03*, Division of Research, College of Business Administration, University of South Carolina, 1995.

Cummings, Ronald G.; Harrison, Glenn W., and Rutström, E. Elisabet, "Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive Compatible?" *American Economic Review*, 85(1), March 1995, 260-266.

Cummings, Ronald G. and Taylor, Laura O., "Does Realism Matter in Contingent Valuation Surveys?" *Land Economics*, 74[2), 1998, 203-215.

de-Magistris, Tiziana; Gracia, Azucena; and Nayga Jr., Rodolfo M., "On the Use of Honesty Priming Tasks to Mitigate Hypothetical Bias in Choice Experiments," *American Journal of Agricultural Economics*, 95(5), August 2013, 1136-1154.

Department of the Interior, "Proposed Rules for Valuing Environmental Damages," *Federal Register*, 59(85), May 4, 1994, 23098-23111.

Fifer, Simon; Greaves, Stephen, and Rose, John, "Hypothetical bias in Stated Choice Experiments: Is it a problem? And if so, how do we deal with it?" *Transportation Research Part A*, 61, 2014, 164-177.

Fifer, Simon; Greaves, Stephen; Rose, John; and Ellison, Richard, "A Combined GPS/Stated Choice Experiment to Estimate Values of Crash-Risk Reduction," *Journal of Choice Modelling*, 4(1), 2010, 44-61.

Fox, John A.; Shogren, Jason F.; Hayes, Dermot J., and Kliebenstein, James B., "CVM-X: Calibrating Contingent Values with Experimental Auction Markets," *American Journal of Agricultural Economics*, 80, August 1998, 455-465.

Gao, Xiaoxue Sherry; Harrison, Glenn W., and Tchernis, Rusty, "Behavioral Welfare Economics and Risk Preferences: A Bayesian Approach," *Experimental Economics*, 25, 2022, forthcoming.

Gibbard, A., "Manipulation of Voting Schemes: A General Result," *Econometrica*, 41, 1973, 587-601.

Haghani, Milad; Bliemer, Michael C.J.; Rose, John M.; Oppewal, Harmen, and Lancsar, Emily, "Hypothetical bias in stated choice experiments: Part I. Macro-scale analysis of literature and integrative synthesis of empirical evidence from applied economics, experimental psychology and neuroimaging," *Journal of Choice Modeling*, 41, 2021a, 100309, https://doi.org/10.1016/j.jocm.2021.100309

Haghani, Milad; Bliemer, Michael C.J.; Rose, John M.; Oppewal, Harmen, and Lancsar, Emily, "Hypothetical bias in stated choice experiments: Part II. Conceptualisation of external validity, sources and explanations of bias and effectiveness of mitigation methods," *Journal of Choice Modeling*, 41, 2021b, 100322, https://doi.org/10.1016/j.jocm.2021.100322.

Harrison, Glenn W, "Theory and Misbehavior of First-Price Auctions," *American Economic Review*, 79, September 1989, 749-762.

Harrison, Glenn W., "Theory and Misbehavior of First-Price Auctions: Reply," *American Economic Review*, 82, December 1992, 1426-1443.

Harrison, Glenn W., "Hypothetical Bias Over Uncertain Outcomes," in J.A. List (ed)., *Using Experimental Methods in Environmental and Resource Economics* (Northampton, MA: Elgar, 2005).

Harrison, Glenn W., "Making Choice Studies Incentive Compatible," in B. Kanninen (ed)., *Valuing Environmental Amenities Using Stated Choice Studies: A Common Sense Guide to Theory and Practice* (Boston: Kluwer, 2006a, 65-108).

Harrison, Glenn W., "Experimental Evidence on Alternative Environmental Valuation Methods" *Environmental and Resource Economics*, 34, 2006b, 125-162.

Harrison, Glenn W. and Hirshleifer, Jack, "An Experimental Evaluation of Weakest-Link/Best-Shot Models of Public Goods," *Journal of Political Economy*, 97, February 1989, 201-225.

Harrison, Glenn W.; Lau, Morten I., and Yoo, Hong Il, "Risk Attitudes, Sample Selection and Attrition in a Longitudinal Field Experiment," *Review of Economics and Statistics*, 102(3), 2020, 552-568.

Harrison, Glenn W.; Lau, Morten I., and Williams, Melonie B., "Estimating Individual Discount Rates in Denmark: A Field Experiment," *American Economic Review*, 92(5), December 2002, 1606-1617.

Harrison, Glenn W, Martínez-Correa, Jimmy, and Swarthout, J. Todd, "Eliciting Subjective Probabilities with Binary Lotteries," *Journal of Economic Behavior & Organization*, 101, 2014, 128-140.

Harrison, Glenn W., Martínez-Correa, Jimmy; Swarthout, J. Todd, and Ulm, Eric R., "Eliciting Subjective Probability Distributions with Binary Lotteries," *Economics Letters*, 127, 2015, 68-71.

Harrison, Glenn W.; Martínez-Correa, Jimmy; Swarthout, J. Todd, and Ulm, Eric "Scoring Rules for Subjective Probability Distributions," *Journal of Economic Behavior & Organization*, 134, 2017, 430-448.

Harrison, Glenn W., and Rutström, E. Elisabet, "Eliciting Subjective Beliefs About Mortality Risk Orderings," *Environmental & Resource Economics*, 33, 2006a, 325-346.

Harrison, Glenn W., and Rutström, E. Elisabet, "Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods," in C.R. Plott and V.L. Smith (eds.), *Handbook of Experimental Economics Results* (Amsterdam: North-Holland: 2006b).

Harrison, Glenn W., and Rutström, E. Elisabet, "Risk Aversion in the Laboratory," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Harrison, Glenn W., and Swarthout, J. Todd, "Eye-Tracking and Economic Theories of Choice Under Risk," *Journal of the Economic Science Association*, 5(1), August 2019, 26-37.

Harrison, Glenn W., and Swarthout, J. Todd, "Cumulative Prospect Theory in the Laboratory: A Reconsideration," in G.W. Harrison and D. Ross (eds.), *Models of Risk Preferences: Descriptive and Normative Challenges* (Bingley, UK: Emerald, Research in Experimental Economics, 2022).

Hensher, David; Louviere, Jordan, and Swait, Joffre D., "Combining Sources of Preference Data," *Journal of Econometrics*, 89, 1999, 197-221.

Hensher, David A.; Rose, Adam M., and Greene, William H., *Applied Choice Analysis* (New York: Cambridge University Press; Second Edition, 2015).

Hess, Stephane; Rose, John M., and Hensher, David A, "Asymmetric Preference Formation in Willingness to Pay Estimates in Discrete Choice Models," *Transportation Research Part E*, 44, 2008, 847-863 .

Hey, John D., "Experimental Investigations of Errors in Decision Making Under Risk," *European Economic Review*, 39, 1995, 633-640.

Holt, Charles A., and Laury, Susan K., "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), December 2002, 1644-1655.

Holt, Charles A., and Laury, Susan K., "Risk Aversion and Incentive Effects: New Data Without Order Effects," *American Economic Review*, 95(3), June 2005, 902-912.

Jacquemet, Nicolas; Joule, Robert-Vincent; Luchini, Stéphane, and Shogren, Jason F., "Preference Elicitation Under Oath," *Journal of Environmental Economics and Management*, 65(1), 2013, 110-132.

Johannesson, Magnus; Blomquist, Glenn C.; Blumenschein, Karen; Johansson, Per-Olov; Liljas, Bengt, and O'Conner, Richard M., "Calibrating Hypothetical Willingness to Pay Responses," *Journal of Risk and Uncertainty*, 8, 1999, 21-32.

Laury, Susan K.; McInnes, Melayne Morgan, and Swarthout, J. Todd, "Avoiding the Curves: Direct Elicitation of Time Preferences," *Journal of Risk and Uncertainty*, 44(3), June 2012, 181-217.

Ledyard, John O., "Public Goods: A Survey of Experimental Research," in J. Kagel and A.E. Roth (eds.), *The Handbook of Experimental Economics* Princeton, NJ: Princeton University Press, 1995.

List, John A., "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards," *American Economic Review*, 91(5), December 2001, 1498-1507.

List, John A., and Gallet, Craig A., "What Experimental Protocol Influences Disparities Between Actual and Hypothetical Stated Values?" *Environmental and Resource Economics*, 20, 2001, 241-254.

List, John A., and Shogren, Jason F., "Calibration of the Differences Between Actual and Hypothetical Valuations in a Field Experiment," *Journal of Economic Behavior and Organization*, 37, November 1998, 193-205.

List, John A., and Shogren, Jason F., "Calibration of Willingness-to-Accept," *Journal of Environmental Economics and Management*, 43[2), 2002, 219-233.

List, John A.; Sinha, Paramita, and Taylor, Michael, "Using Choice Experiments to Value Non-Market Goods and Services: Evidence from the Field," *Advances in Economic Analysis and Policy*, 6[2), 2006, Article 2; http://www.bepress.com/bejeap/advances/vol6/iss2/art2, accessed January 6, 2014.

Loomis, John, "What's To Know About Hypothetical Bias in Stated Preference Valuartion Studies," *Journal of Economic Surveys*, 25[2), 2011, 363–370.

Louviere, Jordan J.; Hensher, David A., and Swait, Joffre D., *Stated Choice Methods: Analysis and Application* (New York: Cambridge University Press, 2000).

Lusk, Jayson L., and Schroeder, Ted C., "Are Choice Experiments Incentive Compatible? A Test with Quality Differentiated Beef Steaks," *American Journal of Agricultural Economics*, 86[2), may 2004, 467-482.

McElreath, Richard, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Boca Raton, FL: Chapman and Hall/CRC, Second Edition, 2020).

McElreath, Richard, and Smaldino, Paul E. "Replication, Communication, and the Population Dynamics of Scientific Discovery," *PLoS ONE*, 10(8), 2015, e0136088. https://doi.org/10.1371/journal.pone.0136088

Mitchell, Robert C., and Carson, Richard T., *Using Surveys to Value Public Goods: The Contingent Valuation Method* (Baltimore: Johns Hopkins Press, 1989).

Moser, Riccarda, and Raffaelli, Roberta, "Does attribute cut-off elicitation affect choice consistency? Contrasting hypothetical and real-money choice experiments," *Journal of Choice Modelling*, 11, June 2014, 16-29.

Moulin, Hervè, *Axioms of Cooperative Decision Making* (New York: Cambridge University Press, 1988).

Murphy, James J.; Allen, P. Geoffrey; Stevens, Thomas H., and Weatherhead, Darryl, "A Meta-analysis of Hypothetical Bias in Stated Preference Valuation," *Environmental and Resource Economics*, 30, 2005, 313–325.

Nape, Steven W.; Frykblom, Peter; Harrison, Glenn W., and Lesley, James C., "Hypothetical Bias and Willingness to Accept," *Economic Letters*, 78(3), March 2003, 423-430.

Nielsen, Otto Anker, "Behavioral Responses to Road Pricing Schemes: Description of the Danish AKTA Experiment," *Intelligent Transportation Systems*, 8, 2004, 233-251.

National Oceanographic and Atmospheric Administration, "Proposed Rules for Valuing Environmental Damages," *Federal Register*, 59(5), January 7, 1994, 1062-1191.

Özdermir, Semra; Johnson, F. Reed; and Hauber, A. Brett, "Hypothetical bias, cheap talk, and stated willingness to pay for health care," *Journal of Health Economics*, 28(4), July 2009, 894-901.

Rasmussen, Eric, *Games and Information: An Introduction to Game Theory* (New York: Basil Blackwell, 1989).

Rossi, Peter E.; Allenby, Greg, M., and McCulloch, Robert, *Bayesian Statistics and Marketing* (Chichester, UK: Wiley, 2005).

Rutström, E. Elisabet, "Home-grown Values and Incentive Compatible Auction Design," *International Journal of Game Theory*, 27, 1998, 427-441.

Samuelson, Paul A., "A Note on the Pure Theory of Consumer's Behavior," *Economica*, 5(17), February 1938, 61–71.

Satterthwaite, M.A., "Strategy-proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory*, 10, 1975, 187-217.

Smaldino, Paul E., and McElreath, Richard, "The Natural Selection of Bad Science," *Royal Society Open Science*, 3, 2016, 160384. http://dx.doi.org/10.1098/rsos.160384.

Smith, Vernon L., "Microeconomic Systems as an Experimental Science," *American Economic Review*, 72(5), December 1982, 923-955.

Svenningsen, Lea S., and Jacobsen, Jette Bredahl, "Testing the effect of changes in elicitation format, payment vehicle and bid range on the hypothetical vias for moral goods," *Journal of Choice Modelling*, 29, December 2018, 17-32.

Swait, Joffre D., and Louviere, Jordan J., "The role of the scale parameter in the estimation and comparison of multinomial logit models," *Journal of Marketing Research*, 30(3), 1993, 305-514.

Varian, Hal .R., "The nonparametric approach to demand analysis," Econometrica, 50(4), 1982, 945–74.

Varian, Hal R., "Non-parametric tests of consumer behaviour," Review of Economic Studies, 50(1), 1993, 99–110.

Vondolia, Godwin K., and Navrud, Ståle, "Are non-monetary payment modes more uncertain for stated preference elicitation in developing countries?" *Journal of Choice Modelling*, 30, March 2019, 73-87.